The 2002 NIST RT Evaluation Speech-to-Text Results

Audrey Le
for the NIST Gang

May 7-8, 2002
Rich Transcription Workshop

Speech-to-Text (STT) Evaluation

- Task:
  - Output orthographic transcript of speech in Hub-4/Hub-5 style
- Scoring:
  - Transcription normalization
  - Overlapping speech was not evaluated
  - For broadcast news, commercials and sports were not evaluated
  - SCLITE software was used to score the results
  - Metrics:
    - Word Error Rate (WER)
    - Normalized Cross Entropy (NCE) for confidence score
    - Matched Pair Sentence Segment Word Error (MAPSSWE) Significance Test

STT Participants and Test Conditions

Domains	Broadcast News		Switchboard				Meetings	
Channels	1		2				Omni	Personal
Speeds	<=10xRT	<=1xRT	>10xRT	<=10xRT		<=1xRT	>10xRT	
Segmentations	Auto	Auto	Manual	Auto	Manual	Manual	Auto	Manual
Sites								
AT&T			P			P		
BBN			P					
CU-HTK			PCC		P			
JHU			PC					
LIMSI	P		PC					
Panasonic	P	P		P	P			
SRI			PC				P	P

P=Primary system submitted
C=Contrastive system submitted

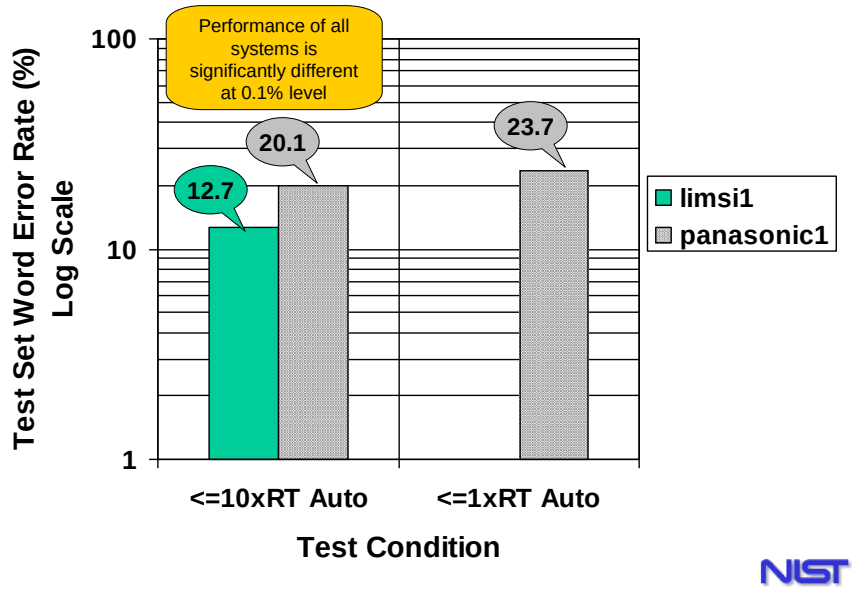
NIST National Institute of Standards and Technology

Broadcast News Test Set

- Duration: 60 minutes
- Consisted of six ~10-minute excerpts from selected shows
  - Selected shows were drawn from the TDT3 data, one show per source, different day of the week
    - MS-NBC
    - PRI the World
    - NBC Nightly News
    - CNN Headline News
    - Voice of America
    - ABC World News Tonight
  - 10-minute randomly-selected excerpts were chosen at story boundaries (one per show)
- Not classified by focus conditions
- Entire broadcasts could be used for unsupervised adaptation
- Cross-broadcast adaptation was not permitted

NIST National Institute of Standards and Technology

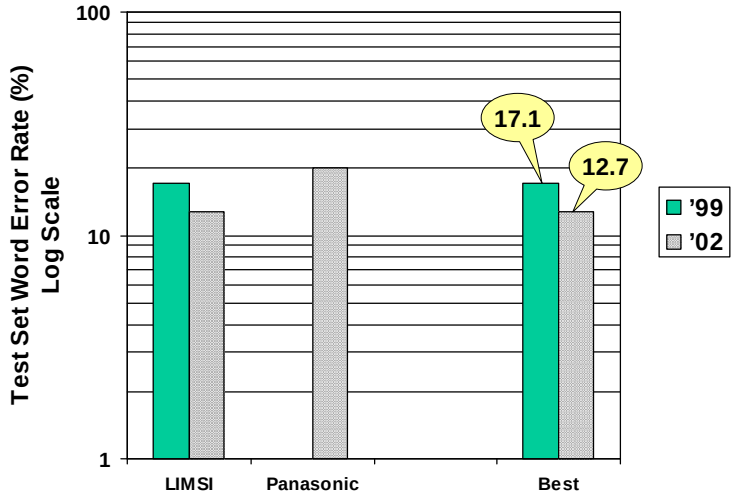
STT Performance on Broadcast News



NIST National Institute of Standards and Technology

Broadcast News Performance History

Speed <= 10xRT, Auto Segmentation



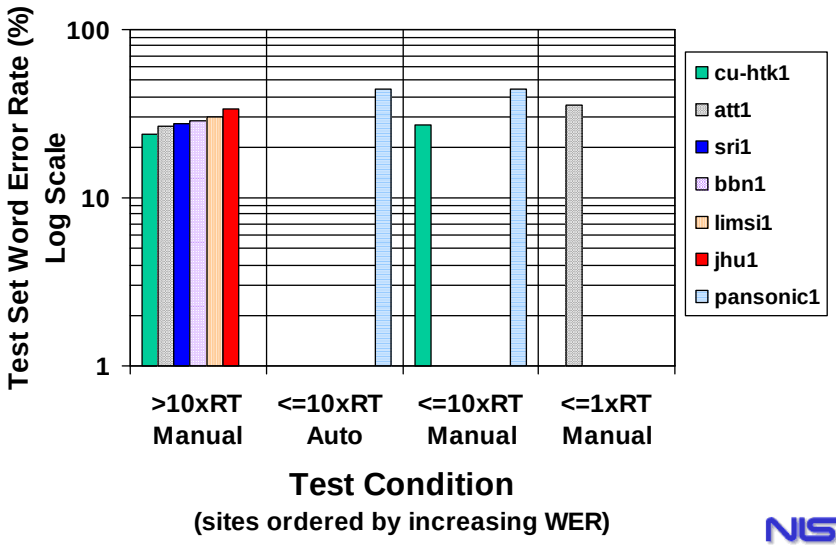
NIST National Institute of Standards and Technology

Switchboard Test Set

- Duration: 300 minutes
- Consisted of sixty ~5-minute conversations from three sources:
  - Twenty conversations from unreleased original SWBD
    - Chosen with least-represented speakers in released corpus
  - Twenty conversations from SWBD II phase 3
    - Selected by LDC
  - Twenty conversations from SWBD Cellular phase 2
    - Chosen to balance gender and call location

NIST National Institute of Standards and Technology

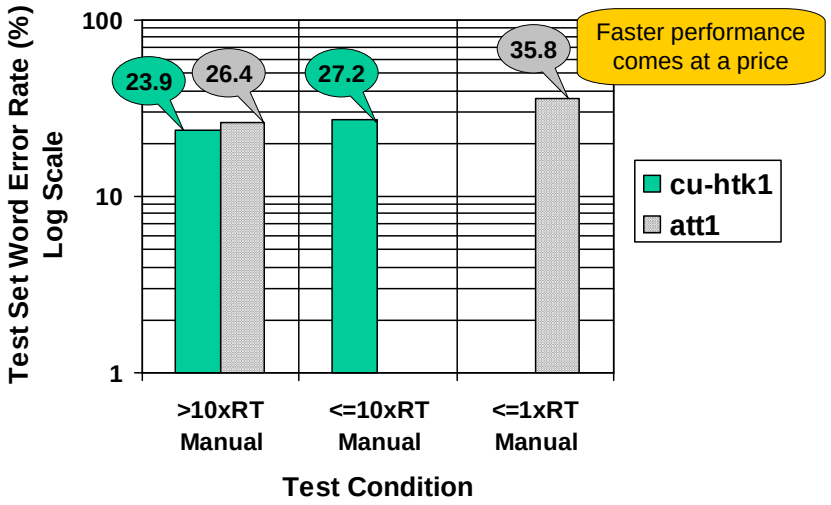
STT Performance on Switchboard



NIST National Institute of Standards and Technology

Speed vs. Accuracy

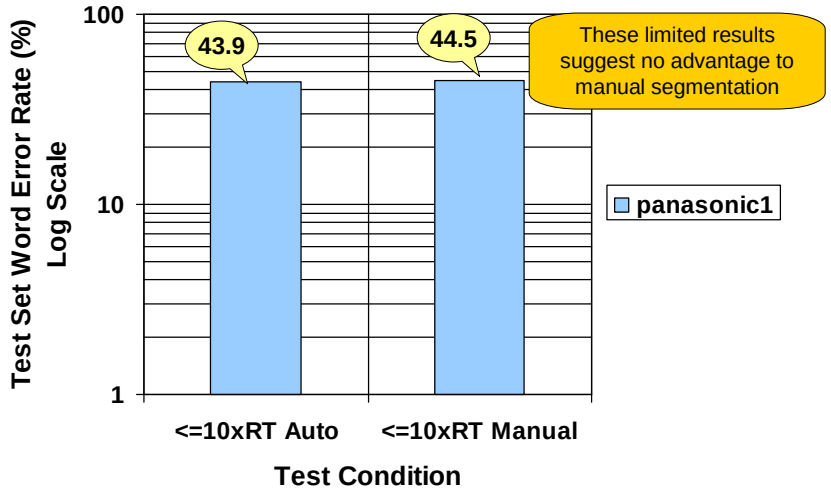
Switchboard Data, Manual Segmentation



NIST National Institute of Standards and Technology

Auto vs. Manual Segmentation

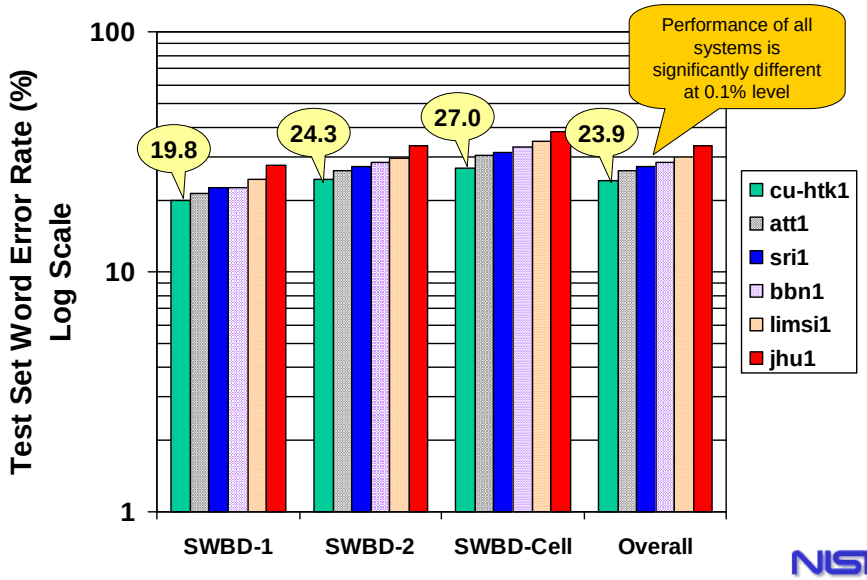
Switchboard Data, Speed <= 10xRT



NIST National Institute of Standards and Technology

STT Performance on Switchboard by Subsets

Speed > 10xRT, Manual Segmentation



NIST National Institute of Standards and Technology

Confidence Metric: Normalized Cross-Entropy (NCE)

$$\text{NCE} = \frac{\left\{ H_{\max} + \sum_{\text{correct}} \log_2(\hat{p}(w)) + \sum_{\text{incorrect}} \log_2(1-\hat{p}(w)) \right\}}{H_{\max}}$$

where $H_{\max} = -n\log_2(p_c) - (N-n)\log_2(1-p_c)$

n = The number of correct HYP words

N = The total number of HYP words

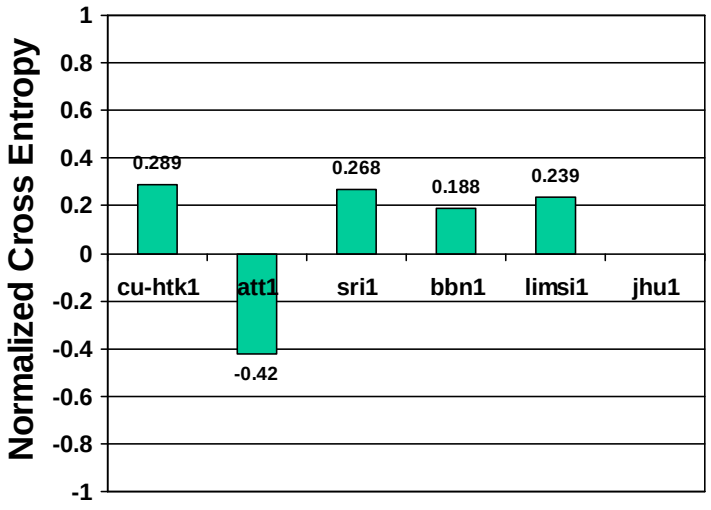
p_c = The average probability that an output word is correct

$\hat{p}(w)$ = The confidence measure output as a function of the output word ***w***

NIST National Institute of Standards and Technology

Confidence Scores on Switchboard Data

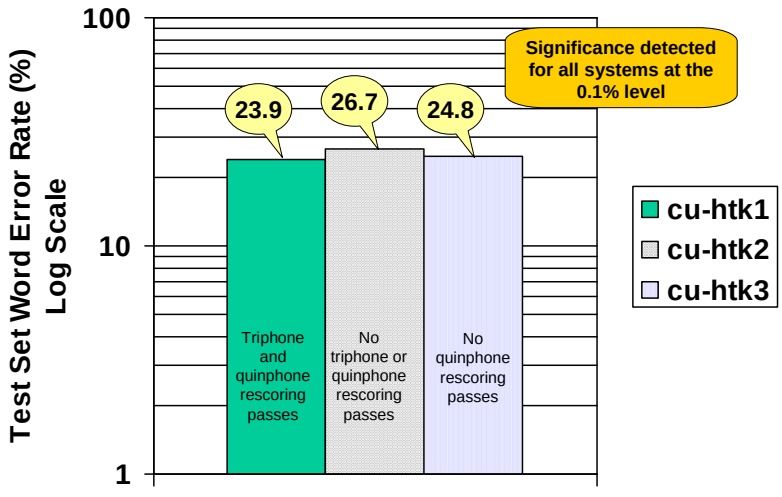
Speed > 10xRT, Manual Segmentation



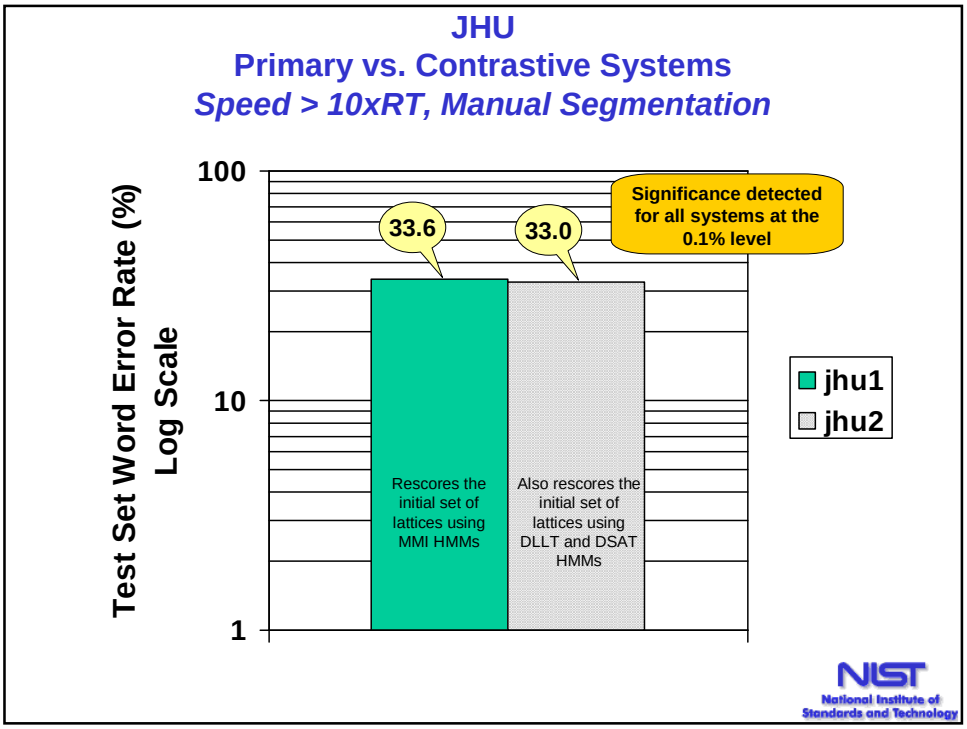
NIST National Institute of Standards and Technology

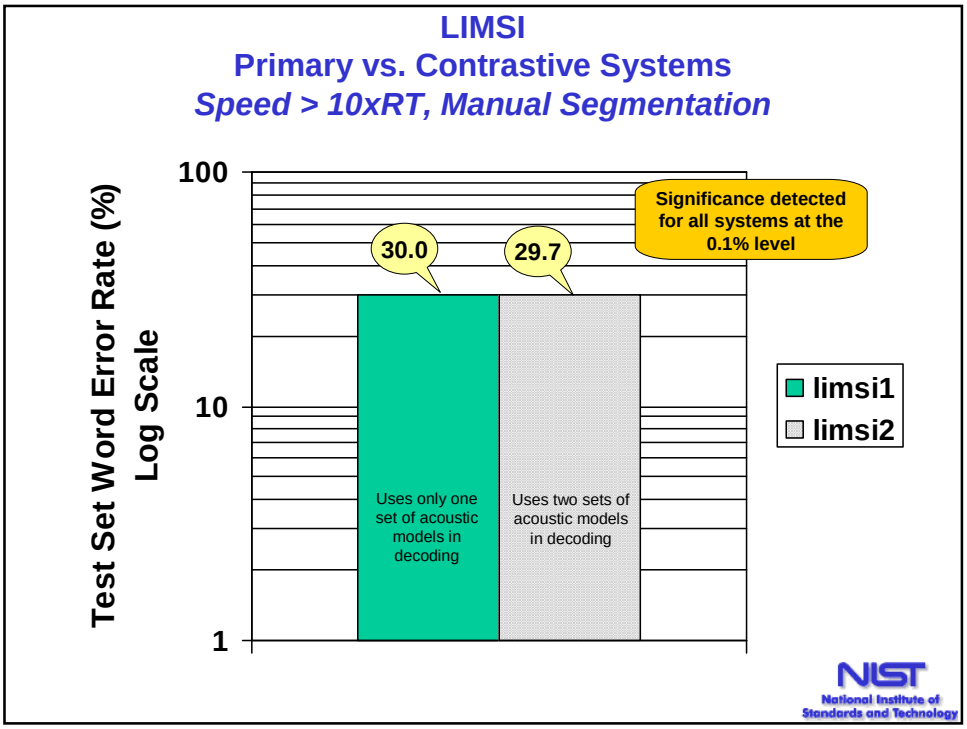
CU-HTK
Primary vs. Contrastive Systems

Speed > 10xRT, Manual Segmentation

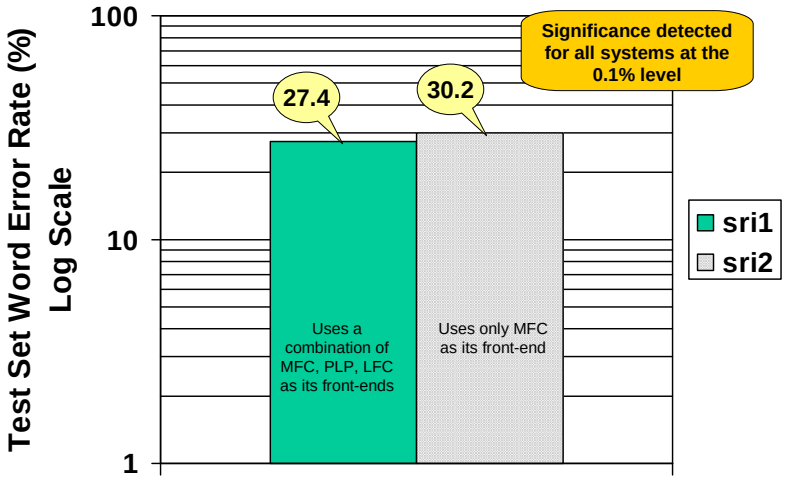


NIST National Institute of Standards and Technology


JHU
Primary vs. Contrastive Systems
Speed > 10xRT, Manual Segmentation
Test Set Word Error Rate (%)
Log Scale
100
10
1
33.6
33.0
Significance detected for all systems at the 0.1% level
Rescores the initial set of lattices using MMI HMMs
Also rescores the initial set of lattices using DLLT and DSAT HMMs
jhu1
jhu2
NIST
National Institute of Standards and Technology

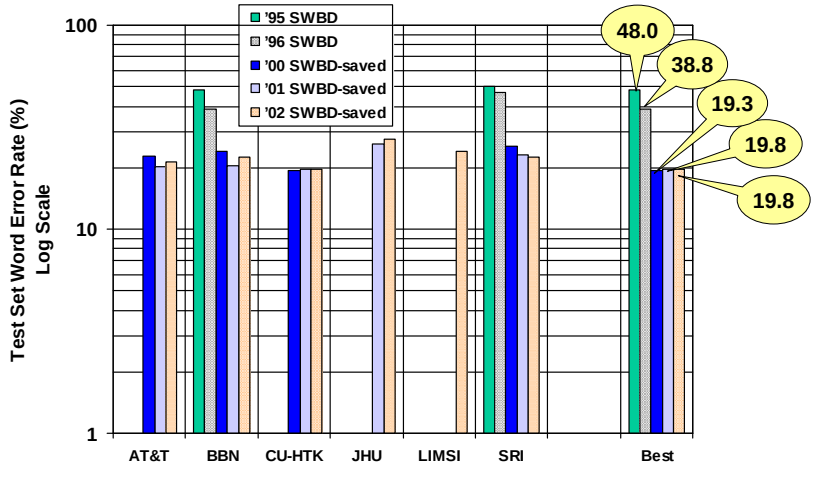

LIMSI
Primary vs. Contrastive Systems
Speed > 10xRT, Manual Segmentation
Test Set Word Error Rate (%)
Log Scale
100
10
1
30.0
29.7
Significance detected for all systems at the 0.1% level
Uses only one set of acoustic models in decoding
Uses two sets of acoustic models in decoding
limsi1
limsi2
NIST
National Institute of Standards and Technology

SRI
Primary vs. Contrastive Systems
Speed > 10xRT, Manual Segmentation



NIST National Institute of Standards and Technology

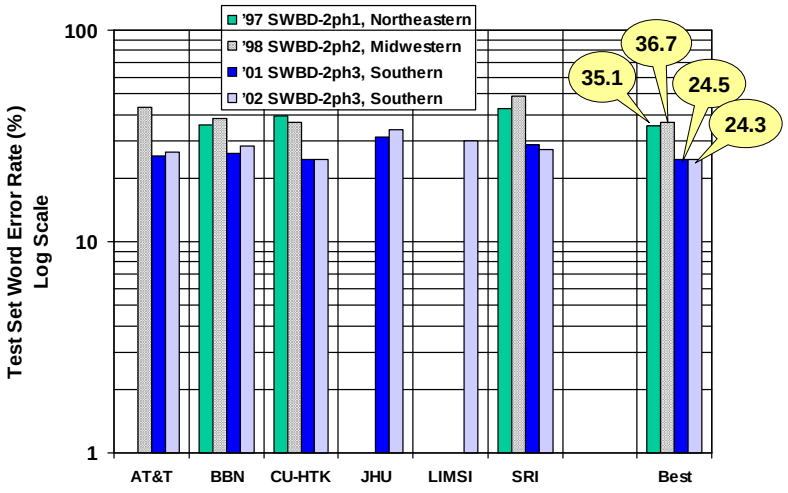
SWBD-1 Performance History
Speed > 10xRT, Manual Segmentation



NIST National Institute of Standards and Technology

SWBD-2 Performance History

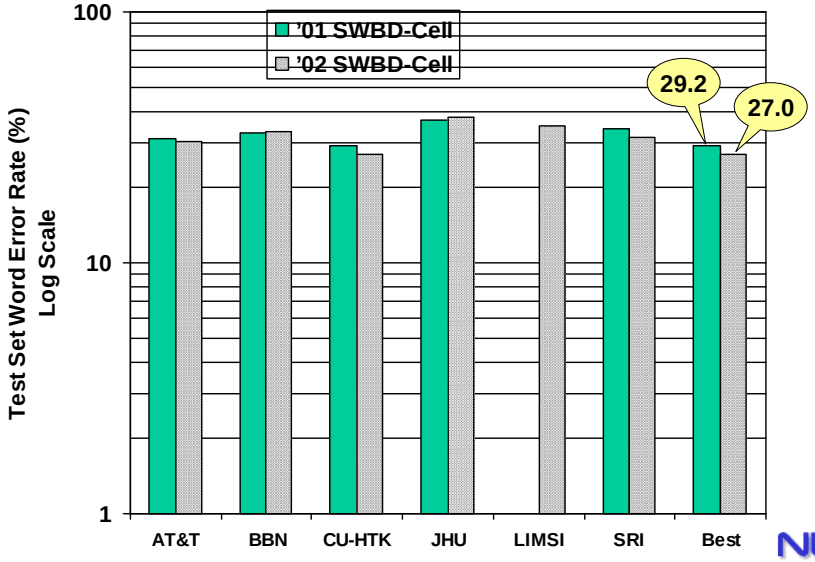
Speed > 10xRT, Manual Segmentation



NIST National Institute of Standards and Technology

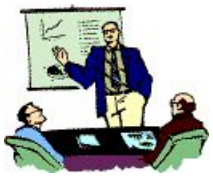
SWBD-Cellular Performance History

Speed > 10xRT, Manual Segmentation



NIST National Institute of Standards and Technology

Meetings Test Set

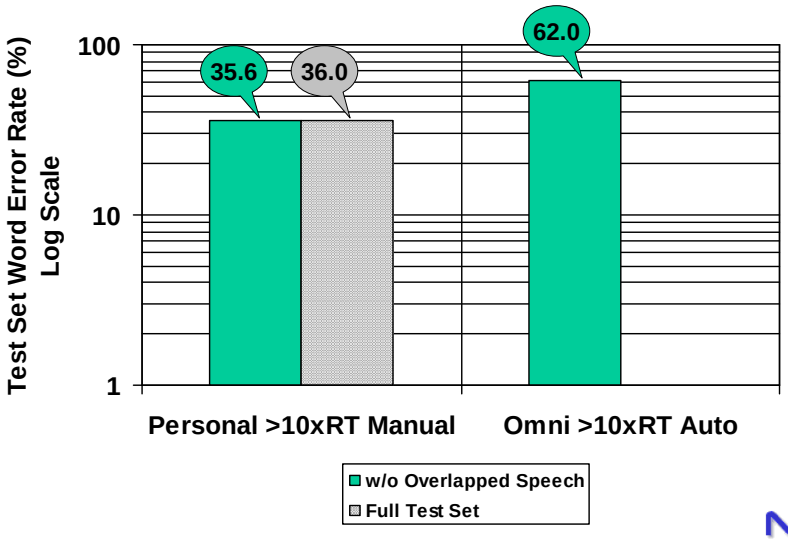


- Duration: 80 minutes
- Consisted of eight ~10-minute excerpts from meetings held at CMU, ICSI, LDC, and NIST
  - Two excerpts from each collection site
- Consisted of recordings from personal, summed personal, and omni microphones
- Entire meetings could be used for unsupervised adaptation
- Cross-meeting adaptation was not permitted

NIST National Institute of Standards and Technology

STT Performance on Meeting Data

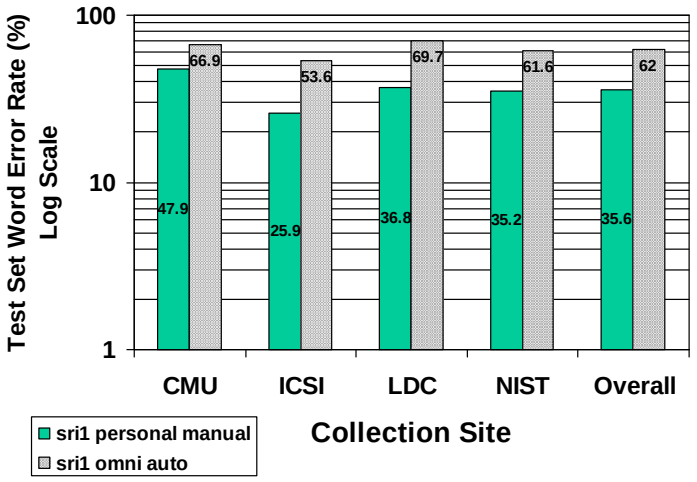
SRI Systems



NIST National Institute of Standards and Technology

STT Performance on Meeting Data by Collection Site

SRI Systems



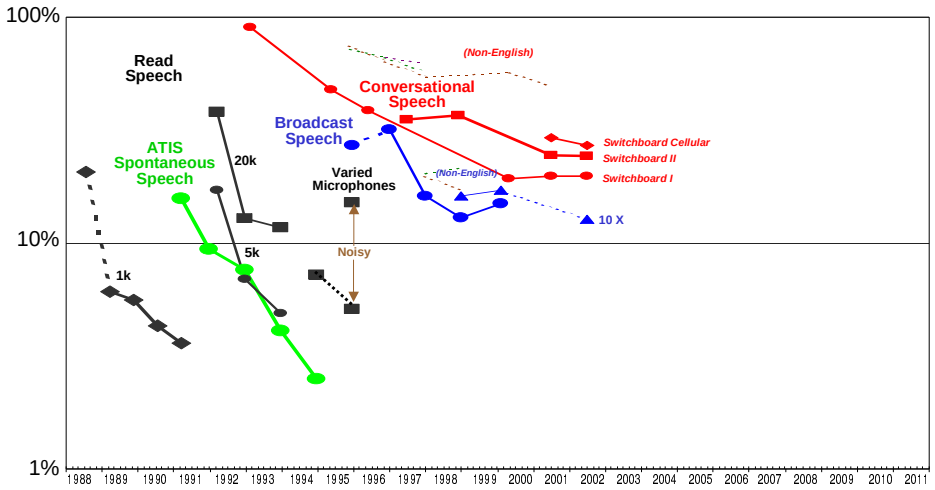
NIST National Institute of Standards and Technology

STT Evaluation Issues

- Test-set to test-set variability still unquantified
  - Bigger test sets?
  - Mothballed systems?
- Overlapping speech a real problem, need to address
- Should NIST provide any speaker segmentation for each domain next year?
- Domain-specific issues
  - Broadcast news
    - Should we measure focus conditions?
  - Telephone Conversations?
  - Meetings
    - Should we continue to measure 3 mic conditions?
      - Omni (primary), Personal, and Personal summed (controls)
    - Should we continue with multi-site test sets?
      - Should there be some/any common constraints on meeting setups? (e.g., mic types, scenarios, AV equipment, etc.)

NIST National Institute of Standards and Technology

NIST STT Benchmark Test History



NIST
National Institute of Standards and Technology